Yoel Melul – 931177240

Machine Learning – Homework 3

Question 1 – Clustering

a. When using the K-means algorithm we minimize the Euclidian metric and thus in the end we find the mean of each of the classes (we will prove this in section b). In contrast when using K-medoid we minimize the L_1 distances to points that are part of our data and thus in the end we find the median of each of the classes (we will prove this in the bonus section). The mean of a set of points is always affected by outliers or noise, while the median is more robust to them. Thus the **K-medoid algorithm is more robust** to noise or outliers than the K-means algorithm.

b. We will prove that the mean (μ) of a set of points x gives the minimum value to the variance of these points, which is equal to the term $\sum_{i=0}^{m}(x_i-\mu)^2$.
We will demonstrate this by proving that for any $a \in \Re$, the minimum of the term $\sum_{i=1}^{m}(x_i-\mu+a)^2$ (where μ is the mean of x) is obtained when $a=0$.
The mean μ is defined:

$$\mu = \frac{1}{m}\sum_{i=1}^{m}(x_i)$$

$$\sum_{i=1}^{m}(x_i-\mu+a)^2 = \sum_{i=1}^{m}[(x_i-\mu)^2+a^2+2a(x_i-\mu)] =$$

$$= \sum_{i=1}^{m}(x_i-\mu)^2 + 2a\sum_{i=1}^{m}(x_i-\mu) + ma^2 =$$

$$= \sum_{i=1}^{m}(x_i-\mu)^2 + 2a\sum_{i=1}^{m}(x_i) - 2am\mu + ma^2 =$$

$$= \sum_{i=1}^{m}(x_i-\mu)^2 + 2am\mu - 2am\mu + ma^2 = \sum_{i=1}^{m}(x_i-\mu)^2 + ma^2$$

m is the quantity of examples in x, the quantity of examples which belong to the class, thus it is a positive number. Then ma^2 is a positive number and $\sum_{i=1}^{m}(x_i - \mu + a)^2$ will be minimal for $a = 0$.

For any $a \in \Re \sim\{0\}$ the term $\sum_{i=1}^{m}(x_i - \mu + a)^2$ will be bigger than the term $\sum_{i=1}^{m}(x_i - \mu)^2$, where μ is the mean of x. Thus, for the 1D case of K-means, the centroid which minimizes the term for variance is the mean of the m examples.

c. Bonus

We will order our data points such that: $x_1 \leq x_2 \leq x_3 \leq \cdots \leq x_m$

The number μ needs to be one of our data points. $\mu \in \{x_1, x_2, \dots, x_m\}$

We will guess than μ is the median of our data points and that we have an odd number of points ($m\ odd$). Then:

$$\mu = x_{\frac{m+1}{2}}$$

We can observe that:

$$\sum_{i=1}^{m}|x_i - \mu| = \left|x_1 - x_{\frac{m+1}{2}}\right| + \left|x_2 - x_{\frac{m+1}{2}}\right| + \cdots + \left|x_m - x_{\frac{m+1}{2}}\right|$$

Be $c \in \left\{\frac{(m+1)}{2} + 1,\ \frac{(m+1)}{2} + 2, \dots, m\right\}$, then:

$$\sum_{i=1}^{m}|x_i - x_c| = |x_1 - x_c| + |x_2 - x_c| + \cdots + |x_m - x_c| =$$

$$= \left|x_1 - x_{\frac{m+1}{2}} - \left(x_c - x_{\frac{m+1}{2}}\right)\right| + \left|x_2 - x_{\frac{m+1}{2}} - \left(x_c - x_{\frac{m+1}{2}}\right)\right| + \cdots$$

$$+ \left|x_m - x_{\frac{m+1}{2}} - \left(x_c - x_{\frac{m+1}{2}}\right)\right|$$

The term $-\left(x_c - x_{\frac{m+1}{2}}\right)$ is negative, and all the terms

$\left(x_1 - x_{\frac{m+1}{2}}\right)$, $\left(x_2 - x_{\frac{m+1}{2}}\right)$, …, $\left(x_{\frac{m+1}{2}-1} - x_{\frac{m+1}{2}}\right)$ are negative too.

In addition, the term $\left(x_{\frac{m+1}{2}} - x_{\frac{m+1}{2}}\right)$ is 0.

The terms $\left(x_{\frac{m+1}{2}+1} - x_{\frac{m+1}{2}}\right)$, $\left(x_{\frac{m+1}{2}+2} - x_{\frac{m+1}{2}}\right)$, … ,$\left(x_m - x_{\frac{m+1}{2}}\right)$ are positive.

This means that any $c \in \left\{\frac{(m+1)}{2}+1,\ \frac{(m+1)}{2}+2, \dots, m\right\}$ will increase the value of the first $\frac{m+1}{2}$ members of the sum and decrease the value of the $\frac{m-1}{2}$ other members by $\left(x_c - x_{\frac{m+1}{2}}\right)$. This means:

$$\sum_{i=1}^{m}|x_i - x_c| = \left|x_1 - x_{\frac{m+1}{2}}\right| + \left|x_2 - x_{\frac{m+1}{2}}\right| + \cdots + \left|x_m - x_{\frac{m+1}{2}}\right| + \left(\frac{m+1}{2} - \frac{m-1}{2}\right)\left(x_c - x_{\frac{m+1}{2}}\right) =$$

$$= \left|x_1 - x_{\frac{m+1}{2}}\right| + \left|x_2 - x_{\frac{m+1}{2}}\right| + \cdots + \left|x_m - x_{\frac{m+1}{2}}\right| + \left(x_c - x_{\frac{m+1}{2}}\right) >$$

$$> \left|x_1 - x_{\frac{m+1}{2}}\right| + \left|x_2 - x_{\frac{m+1}{2}}\right| + \cdots + \left|x_m - x_{\frac{m+1}{2}}\right| = \sum_{i=1}^{m}\left|x_i - x_{\frac{m+1}{2}}\right|$$

In a similar way, if $c \in \left\{1,\ 2, \dots, \frac{m+1}{2}-1\right\}$, then the term $-\left(x_c - x_{\frac{m+1}{2}}\right)$ is positive, and in this case the last $\frac{m+1}{2}$ members of the sum will increase and the first $\frac{m-1}{2}$ members will decrease by $\left(x_{\frac{m+1}{2}} - x_c\right)$. This means:

$$\sum_{i=1}^{m}|x_i - x_c| = \left|x_1 - x_{\frac{m+1}{2}}\right| + \left|x_2 - x_{\frac{m+1}{2}}\right| + \cdots + \left|x_m - x_{\frac{m+1}{2}}\right| + \left(\frac{m+1}{2} - \frac{m-1}{2}\right)\left(x_{\frac{m+1}{2}} - x_c\right) =$$

$$= \left|x_1 - x_{\frac{m+1}{2}}\right| + \left|x_2 - x_{\frac{m+1}{2}}\right| + \cdots + \left|x_m - x_{\frac{m+1}{2}}\right| + \left(x_{\frac{m+1}{2}} - x_c\right) >$$

$$> \left|x_1 - x_{\frac{m+1}{2}}\right| + \left|x_2 - x_{\frac{m+1}{2}}\right| + \cdots + \left|x_m - x_{\frac{m+1}{2}}\right| = \sum_{i=1}^{m}\left|x_i - x_{\frac{m+1}{2}}\right|$$

Thus, if $\mu = x_{\frac{m+1}{2}} = median$ the term $\sum_{i=1}^{m}|x_i - \mu|$ will have its minimum value for any $\mu \in \{x_1, x_2, \dots, x_m\}$.

If m was even and $\mu = x_{\frac{m}{2}+1}$, then a similar approach for $c \in \left\{\frac{m}{2}+2,\ \frac{m}{2}+3, \dots, m\right\}$ would lead to the last $\frac{m}{2}-1$ members of the sum to decrease and the first $\frac{m}{2}+1$ to increase by $\left(x_c - x_{\frac{m}{2}+1}\right)$. This means:

$$\sum_{i=1}^{m}|x_i - x_c| = \left|x_1 - x_{\frac{m}{2}+1}\right| + \left|x_2 - x_{\frac{m}{2}+1}\right| + \cdots + \left|x_m - x_{\frac{m}{2}+1}\right| + \left(\frac{m}{2}+1-\left(\frac{m}{2}-1\right)\right)\left(x_c - x_{\frac{m}{2}+1}\right) =$$

$$= \left|x_1 - x_{\frac{m}{2}+1}\right| + \left|x_2 - x_{\frac{m}{2}+1}\right| + \cdots + \left|x_m - x_{\frac{m}{2}+1}\right| + 2\left(x_c - x_{\frac{m}{2}+1}\right) >$$

$$> \left|x_1 - x_{\frac{m}{2}+1}\right| + \left|x_2 - x_{\frac{m}{2}+1}\right| + \cdots + \left|x_m - x_{\frac{m}{2}+1}\right| = \sum_{i=1}^{m}\left|x_i - x_{\frac{m}{2}+1}\right|$$

For $c \in \left\{1,\ 2, \dots, \frac{m}{2}-1\right\}$ would lead to the first $\frac{m}{2}-1$ members of the sum to decrease and the last $\frac{m}{2}+1$ to increase by $\left(x_{\frac{m}{2}+1} - x_c\right)$. This means:

$$\sum_{i=1}^{m}|x_i - x_c| = \left|x_1 - x_{\frac{m}{2}+1}\right| + \left|x_2 - x_{\frac{m}{2}+1}\right| + \cdots + \left|x_m - x_{\frac{m}{2}+1}\right| + \left(\frac{m}{2}+1-\left(\frac{m}{2}-1\right)\right)\left(x_{\frac{m}{2}+1} - x_c\right). =$$

$$= \left|x_1 - x_{\frac{m}{2}+1}\right| + \left|x_2 - x_{\frac{m}{2}+1}\right| + \cdots + \left|x_m - x_{\frac{m}{2}+1}\right| + 2\left(x_{\frac{m}{2}+1} - x_c\right). >$$

$$> \left|x_1 - x_{\frac{m}{2}+1}\right| + \left|x_2 - x_{\frac{m}{2}+1}\right| + \cdots + \left|x_m - x_{\frac{m}{2}+1}\right| = \sum_{i=1}^{m}\left|x_i - x_{\frac{m}{2}+1}\right|$$

If $c = \frac{m}{2}$ then $\frac{m}{2}$ members of the sum would increase and $\frac{m}{2}$ members would decrease, leading to the sum being equal:

$$\sum_{i=1}^{m}\left|x_i - x_{\frac{m}{2}}\right| = \sum_{i=1}^{m}\left|x_i - x_{\frac{m}{2}+1}\right|$$

Thus, if $\mu \in \left\{x_{\frac{m}{2}}, x_{\frac{m}{2}+1}\right\} = median$ the term $\sum_{i=1}^{m}|x_i - \mu|$ will have its minimum value for any $\mu \in \{x_1, x_2, \dots, x_m\}$.

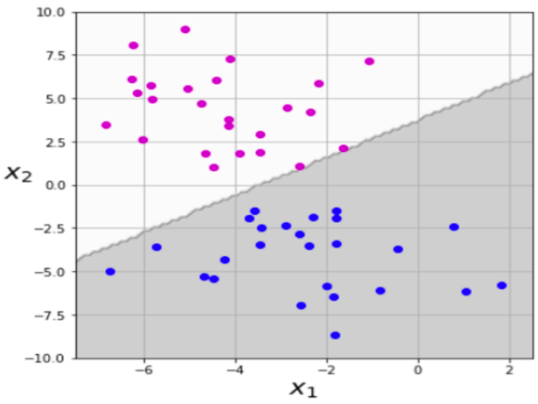
Quot erat demonstrandum.

Question 2 – SVM

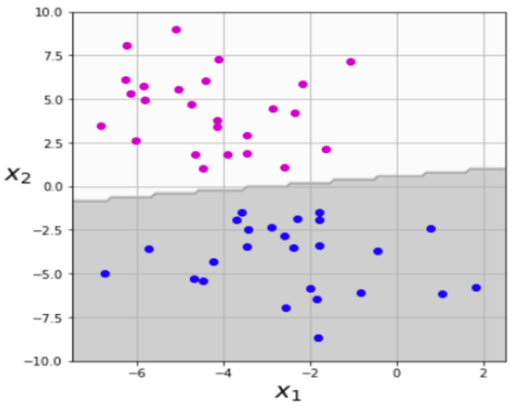
Linear

First, we observe that the only linear SVMs are A and D. Thus, we will match them with the two linear kernels (1 and 2).

A



D



We can observe that in figure A there is a bigger margin to the data of the two classes but there is a bigger risk of misclassification of training data points. The parameter C defines how much penalization to give misclassificated data, thus a bigger C will have lower risk of misclassification of training data rather than a bigger margin.

Thus:

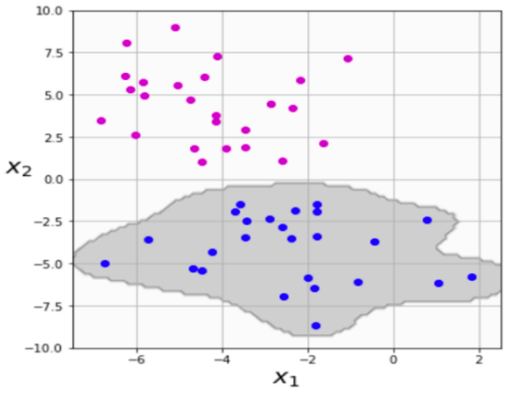
Figure A matches with 1. Linear kernel with $C = 0.01$.

Figure D matches with 2. Linear kernel with $C = 1$.

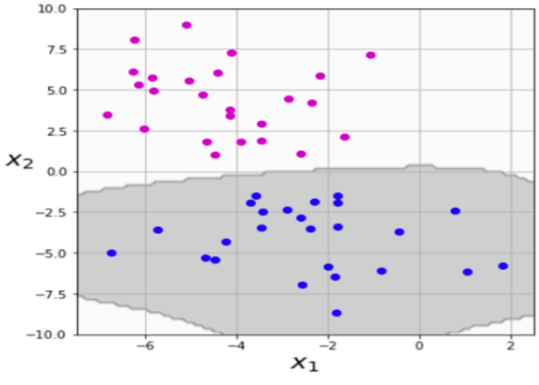
RBF

Next, we observe that in figures B and E, the grey class (to which the blue data belong) is “closed”, surrounding the blue data. This is only possible using RBFs kernel; thus, we will match these figures with the two RBF kernels (5 and 6).

B



E



The parameter γ is related to how much we fit the data. We can observe that figure B is probably overfitting and thus it belongs to the model with bigger γ.

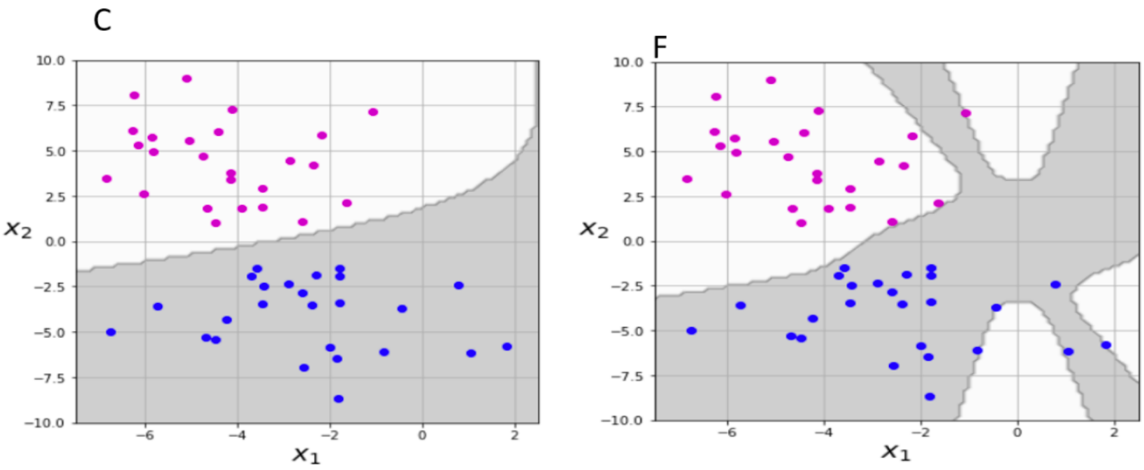
Thus:

Figure B matches with 6. RBF kernel with $\gamma = 1$

Figure E matches with 5. RBF kernel with $\gamma = 0.2$

Polynomial

The two last figures, which belong to polynomial kernels (3 and 4) are figures C and F.



We can observe that in figure F the model is clearly overfitting, which is related to a polynomial of higher order.

Thus:

Figure C matches with 3. 2^{nd} order polynomial kernel.

Figure F matches with 4. 10^{nd} order polynomial kernel.

In conclusion:

Figure	A	B	C	D	E	F
Model	1	6	3	2	5	4

Question 3 – Capability of generalization

a. The term that represents the balance in machine learning is the **generalization** of a model. The generalization refers to the ability of a model to adapt to data that has not been used to train the model or its hyperparameters. This generalization is related to the bias-variance tradeoff: if the model fits too well the training data (overfitting) the model might be too complex, with low bias but high variance, that will lead to poor generalization. In contrast, if the model is too simple (underfitting), it might have low variance but high bias, which will lead to poor performance metrics. Finding the balance between these two concepts is what leads to a good generalization of a well-performing model.

b. As explained by the exercise, p is the total number of learnt parameters. As we increase this number, the model will fit better the training data (decreasing bias), but the complexity will increase, and at some point there will be overfitting (increase in variance). Thus, as the term 2p increases, the bias decreases and the variance increases.

   As explained by the exercise, $\hat{L}$ is the estimated likelihood given p parameters. $\hat{L}$ represents the goodness of fit of our model; as our model fits better the data, $\hat{L}$ will increase. This is why $\hat{L}$ increases as the number of parameters in our model increase. Thus, as the term $\hat{L}$ increases, $-2\ln(\hat{L})$ decreases, the bias decreases and the variance increases.

c. If the balance between the two terms 2p and $-2\ln(\hat{L})$ is violated, the balance between bias and variance will be violated too, causing our model to be underfitting (high bias, low variance) if 2p is low but $-2\ln(\hat{L})$ is high, or overfitting (low bias, high variance) if 2p is high but $-2\ln(\hat{L})$ is low.

d. With the Akaike Information Criterion (AIC) we are aiming to keep the balance between bias and variance in order to have good generalization. The ideal AIC is as minimum as possible. This means the complexity of our model (term 2p) matches

the goodness of fit (term $-2\ln(\hat{L})$). As our model increases in complexity the AIC is penalized by the term 2p, but if this increase in complexity led to a significant increase in goodness of fit then the term $-2\ln(\hat{L})$ will lower the AIC enough and surpass the penalization of 2p. This is how AIC helps us reach a good balance of the bias and variance.